基于词向量的汉语成语的语义透明度分析

摘要：本文基于语义透明度的计算公式，计算了部分汉语成语的语义透明度。基于实验结果，我们发现汉语成语的语义透明度整体偏低，这是因为成语蕴含着丰富的文化意义，若不了解成语背后的文化故事，便很难理解成语的语义，致使其语义透明度低；成语字面义与衍生义之间的关联度也是影响成语语义透明度的关键因素，关联度高则语义透明度也高。语义透明度相对较高的成语多为联合式语法结构，并且成语整体的语义与两个组成部分的语义经常相等或相近。

关键词：成语；语义透明；语法结构

0. 引言

1962年，英国功能派语义学家S.Ullmann在《语义学：意义科学导论》一书中提到“语义透明 / 隐晦的词“的说法。Schreuder&Baayen(1995)对语义透明度进行了诠释：一个复合词与其组成成分之间语义层面上的透明度关系取决于复合词整体及其构成词素语义表征集合之间的重叠程度。“语义透明度”这一理论大概于90年代前后被引进国内，随后很快在语言学、教育学以及心理学等多种领域里均产生了很大的影响。目前国内学界普遍认可的是王春茂、彭聃龄(1999) 提出的关于语义透明度的界定：语义透明度是指复合词的语义可从组成复合词的各个词素的语义推知的程度，其操作定义为整词与其词素的语义相关程度。语义透明通常被分为：透明、比较透明、比较晦涩、完全晦涩四种类型。

语义透明度的研究在儿童阅读和对外汉语等领域都有一定的应用价值，徐彩华（2001）研究了语义透明度对于儿童阅读的影响，认为透明词有利于学习，不透明词对学习造成一定障碍；柳莉（2011）发现在对外汉语教学过程中，透明度高的词语，学生容易把握，透明度较低的词，学生把握起来比较困难，常常有望文生义的情况。

成语是一种相沿习用、含义丰富、具有书面语色彩的固定短语。成语在表意上与一般固定短语不同，它的意义往往并非其构成成分意义的简单相加，而是在其构成成分意义的基础上进一步概括出来的整体意义，具有意义整体性的特点。成语在语文教学以及汉语二语教学中都扮演着重要的角色，学生掌握一定量的成语，有利于写作和言语交际。如何确定哪些成语适合哪一教学阶段，我们需要从多个角度衡量成语的学习难度。本文将从成语的语义透明度着手，计算分析部分汉语成语的语义透明，以期望为汉语成语研究者提供些许参考。

1. 研究现状

chinaXiv:202011.00125v1

chinaXiv:202011.00125v1

在影响词语透明度的因素和语义透明度的影响研究方面，许艳华（2014）利用 spss 方差分析确定影响语义透明度的因素，采用多元回归的方式来计算不同因素的影响程度，给不同的因素赋予不同的权重，从而构造出语义透明度的计算公式；孙威（2016）基于类义词典《同义词词林》，以双音节复合词的词素义作为参考标准,通过计算复合词整词与词素之间的语义距离,评估两者间语义相关程度,构建语义透明度计算公式从而评估划分透明度等级。

在成语语义透明度研究方面，付培丽（2012）以语素义是否常用，成语整体义是否字面义为判定标准，将 469 个成语划分为四个等级。赵丹（2016）从对外汉语教学研究的角度，将学习者对成语语素义的认知难度作为成语语义透明度判定标准，将透明度按完全透明到完全晦涩氛围了四个级别，对不同语言水平学习者的成语语义透明度差异进行了考察。吴迪（2016）考察了语义透明度因素对成语使用率的影响，发现高透明度成语的使用频率高于低透明成语。柴湘露（2018）对比成语义变前后，发现成语义变往往伴随成语字义显著度、内部组合关系显著度、字面义与衍生义关联度的提高，成语语义变化具有语义更透明的倾向。

在英语成语语义透明研究方面，Hee-Rahk Chae（2015）通过分析成语的内部结构，认为英语成语中的形容词没有承担成语的部分意义，所以成语的语义透明度会比较低。Frank Boers(2015)通过比较母语者和二语学习者对成语语义透明度的评级，发现两者的评级之间存在显著的差异。

目前关于成语的语义透明度的研究方法主要是依靠人工评定，效率低下，无法定量批量地分析。本文以汉语成语为研究对象，利用词向量，计算了部分成语的语义透明度，基于实验结果，探究影响汉语成语的语义透明度的因素。

2. 语义透明度计算公式

在之前的研究中，我们已经提出了计算汉语复合词语义透明度的公式。根据语义透明度的的定义，将语义透明度转化为各个词素的语义与整词语义的相似度。如果词素与整词的语义相似度高，那么说明能够从词素推导出复合词语义的程度就高，词素与整词的语义相似度低，那么从词素推导出复合词语义的程度就低。用词向量代表整词的语义，字向量代表词素的语义，通过计算词与字的语义相似度，计算复合词的语义透明度。我们将构词语素字向量相加再取均值，然后和整词的向量做相似度计算，即可以得到整词的语义透明度。我们的语义透明度计算公式如下：

$$c_m = \frac{\sum_{i=1}^{n} c_i}{n} \qquad (1)$$

c_m是词素向量的平均值，c_i为词素向量，n为组成复合词的词素数。

$$s = \frac{w \times c_m}{\sqrt{w^2 + {c_m}^2}} \qquad (2)$$

选用余弦距离计算语义相似度，如函数2 所示，s为词素平均值和复合词整词的相似度，w为整词向量。

$$w_t = s + 0.5 \times s \qquad (3)$$

为了便于分析，如公式（3）我们将（2）得到语义相似度进行归一化处理。w_t即为复合词的语义透明度。

在本文中，我们将计算成语的语义透明度，在计算时，现将成语切成语素。如“老马识途”切为“老”、“马”、“识”、“途”四个语素。“囫囵吞枣”由于“囫囵”切为“囫囵”、“切”、“枣”三个语素，“囫囵”是一个语素，所以不再切分。将切分得到的语素的向量相加，然后取平均，再与整个成语的向量做相似度计算，对结果进行归一化，即得到该成语的语义透明度。

3. 实验

3.1 数据集和实验结果

本文中，我们将经过清洗的 6.2G 维基百科无标注语料作为数据集。选择 word2vector 作为词向量训练工具。word2vector 是 Google2013 年推出的用于获取词向量的工具包，word2vector 可以根据给定的语料库，通过优化后的训练模型快速有效地将一个词语表达成向量形式，本文选用赵哲（2017）提出的 n-gram 方法优化 word2vector 模型训练词向量，这种方法可以通过学习 n-gram 共现信息从而提高词向量的质量，捕捉到更多的语义信息。

3.2 实验结果与分析

随机从《新华成语词典》中抽取 400 个成语，利用第二节介绍的语义透明度计算公式计算成语的语义透明度。结果如表 1。从表 1 所示结果来看，成语的语义透明整体偏低，皆低于 0.6。

chinaXiv:202011.00125v1

NO	成语	透明度	成语	透明度	成语	透明度
1	悬崖峭壁	0.584	相辅相成	0.425	完璧归赵	0.243
2	深情厚谊	0.508	心甘情愿	0.419	小心翼翼	0.237
3	狼吞虎咽	0.482	优胜劣汰	0.399	三顾茅庐	0.231
4	诚心诚意	0.467	轰轰烈烈	0.363	助纣为虐	0.223
5	苦尽甘来	0.449	战战兢兢	0.361	卧薪尝胆	0.205
6	各抒己见	0.445	吞吞吐吐	0.361	破釜沉舟	0.198
7	肆无忌惮	0.442	后顾之忧	0.361	乐不思蜀	0.178
8	飞禽走兽	0.441	名胜古迹	0.360	东施效颦	0.171
9	丢三落四	0.437	日新月异	0.334	望梅止渴	0.134
10	知足常乐	0.436	雪上加霜	0.315	庖丁解牛	0.122

表 1.成语语义透明度部分计算结果

基于实验结果我们将分析影响成语语义透明度的因素。如表1所示，语义透明度较高的“悬崖峭壁”、“深情厚谊”、“狼吞虎咽”等成语皆为联合式语法结构。联合式语法结构一般由相近词性的语素构成，成语一般都两个词语构成，构成联合式成语的两个词语一般都具有相同的语法结构。如构成“丢三落四”的两个词都是动宾结构，为动词性词语；构成“悬崖峭壁”的两个词皆是定中结构，名词性词语。

在这些联合式成语的内部语义关系也具有一些特点，往往成语的语义与组成成语的两个词语具有语义相等或相近的特点。如“深情厚谊”中的“深情”与“厚谊”在语义上都表示友谊深厚；“狼吞虎咽”中的“狼吞”、“虎咽”都表示吃东西又急又忙。由于构成成分的语义相近，从构成部分的语义推知整个成语的意义也就更容易，语义透明度也就更高。“心甘情愿”、“优胜劣汰”都属于此类型的成语。

AABB联合式成语的语义透明度也相对较高。词语重叠是汉语的一种语法手段，词语在重叠之后，语义发生了一些变化。李明宇在《论词语重叠的意义》中提到，重叠的基本语法意义是调量，即量的变化。重叠式成语表示的也是量的增加。如“家家户户”指每一家每一户，所有人家。与重复之前的“家”、“户”相比，语义上数量增加。“形形色色”、“吞吞吐吐”都是此类成语。该类AABB重复联合式成语，其形式和意义都存在量增加，所以由原来的组成成分A、B推知整个成语意义，只用在原来的语义上多一项量增加的语义即可，所以语义透明度相对较高。

前面我们提到成语的语义透明度整体偏低，我们猜想成语的文化意义是导致语义透明度的重要因素之一。很多成语是对古代典籍和历史事件的概括和总结，饱含着丰富的文化意义，其意义很难从字面推知。如表 1 后半段所示，我们计算了部分概括历史事件成语的语义透明度。从表示结果来看，整体结果皆低于 0.3，若不了解这些成语背后的故事，其

chinaXiv:202011.00125v1

语义很难从其组成成分的语义推知。如“三顾茅庐”如果了解刘备访聘诸葛亮的故事，即知该成语“比喻真心诚意，一再邀请”。成语不仅是一种语言现象，也是一种文化现象，成语蕴含的文化意义在一定程度上使得其语义透明度比较低。

同时汉语成语具有表意双层性的特点，刘洁修（1985）将成语的意义分为三类：成语的字面义、成语的引申义和成语的比喻义。字面义即为组成成语的语素意义的直接相加，引申义和比喻义是在字面意义的基础上衍生出的新义。字面义与衍生义的关联度也是影响成语语义透明度的重要因素。

当字面义与衍生义关联度高时，成语的语义透明度相对较高。例如构成“日新月异”的四个语素都为常用语素，意义也比较简单，字面意义的简单相加“每天每月都有新的变化”，常用义项为字面意义引申出来的意义 “形容进步、发展很快”。该“日新月异”常用的义项并虽然非简单的字面意义，但与字面意义的关联度高，并且字面意义是由构成成语语素的常用义项的简单相加。“雪上加霜”的常用义项是其比喻意义“比喻接连遭受灾难，损害愈加严重”，该比喻义由字面义隐喻而来。字面义与衍生义关联度相对比较高，从字面义推知比喻义比较容易，所以语义透明度较高。“小心翼翼”的字面义为“严肃恭敬”，常用义项“现形容谨慎小心，一点不敢疏忽”，字面义与常用义项之间的关联度相对较低，并且“翼翼”在现代汉语中不能单独成词，其意义“严肃恭敬”在现代汉语中少用，所以致使“小心翼翼”的语义透明度低。

综上，经过实验计算我们发现语义透明度相对较高的成语多为联合式语法结构，并且成语整体的语义与两个组成部分的语义经常相等或相近。实验结果表明成语的语义透明度整体偏低，这是因为成语蕴含着丰富的文化意义，若不了解成语背后的文化故事，便很难理解成语的语义，致使其语义透明度低；成语字面义与衍生之间的关联度也是影响成语语义透明度的关键因素，关联度高则语义透明度也高。

4. 应用

经过以上分析，语义透明度较高的成语相对来说，结构简单，语义易从字面推知，在教学过程中，教师可以先教授这部分语义透明度较低的成语，便于学生理解掌握。对于部分语义透明度较低的成语，例如“卧薪尝胆”、“东施效颦”此类包含成语故事的成语，可以伴随着成语故事进行教学，让学生从故事领会成语的含义。

5. 总结

本文基于语义透明度的定义，将语义透明度的计算转化为各个词素的语义与整词语义的相似度的计算。基于词向量的语义透明的计算公式。计算了部分成语的语义透明度，成

chinaXiv:202011.00125v1

语的语义透明度整体偏低，这与成语的表意双层性特点密切相关；汉语成语包含丰富的文化意义也是致使成语透明度偏低的因素。

参考文献

[1] 王春茂,彭聃龄. 合成词加工中的词频、词素频率及语义透明度[J]. 心理学报：1999 年 7 月 31 卷 3 期

[2] 李晋霞,李宇明. 论词义透明度[J].第三期语言研究：2008 年 7 月第 28 卷第三期

[3] 徐彩华,李镗. 语义透明度影响儿童词汇学习的实验研[J]. 语言文字应用:2001 年 01 期

[4] 柳莉. 语义透明度和词的利举行在对外汉语教学中的作用[J].《西南大学学报》:2011 年 12 月第 28 卷第 6 期

[5] 宋宣. 汉语偏正复合名词语义透明的的判定条件[J]. 云南师范大学学:2011 年 03 期

[6] 宋贝贝. 现代汉语动名型复合词词义透明度研究[J]. 语言文字应用:2015 年 8 月第 3 期

[7] 刘伟. 汉语定中式双音节复合词的词典语义透明度研究[J]. 鲁东大学学报:2016 年 5 月第 33 卷第 3 期

[8] 赵丹. 常用成语语义透明度研究[J]. 现代交际:2016 年 11 月

[9] 孙威. 基于《同义词词林》的双音节复合词语义透明的测评分析[J]. 语文学刊:2016 年 11 期

[10] 任敏. 影响现代汉语双音节符合词语义透明度的机制研究[J]. 河北师范大学学报:2012 年 6 月 35 卷第 4 期

[11] 宋宣. 汉语复合词语义透明度的释义模式分析. 云南师范大学学报:2013 年 5 月第 11 卷第 3 期

[12] Bell, M. J., & Schäfer. M. Semantic transparency: challenges for distributional semantics[C]. In A. Herbelot, R. Zamparelli, & G. Boleda (Eds.), *Proceedings of the IWCS 2013 workshop: Towards a formal distributional semantics*. *Potsdam: Association for Computational Linguistics* (pp. 1–10).

[13] Melanie J. Bell,Martin Schäfer . Modelling semantic transparency Morphology[C]. May 2016, Volume 26, Issue 2, pp 157–199

[14] Yoon Kim. 2014. Convolutional neural networks for sentence classification[C]. In *Proceedings of EMNLP 2014*, pages 1746–1751.

[15] Ronan Collobert, Jason Weston, Le ´on Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch[C]. *Journal of Machine Learning Research*, 12:2493–2537.

[16] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation[C]. In *Proceedings of EMNLP 2014*, pages 1532–1543.

[17] Zhao Zhe. Liu Tao. Li Shen. Li Bofang. Du Xiaoyong. *Ngram2vec: Learning Improved Word Representations from N-gram Co-occurrence Statistics[C]*. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

[18] R.Schreuder. H.Baayen. Modeling morphological processing .In L.B.Feldman (Ed.), Morphological Aspects of Language Processing[C]. Hillsdale, NJ: Lawrence Erlbaum,1995

[19] 孙威. 代汉语并列式双音节符合词的语义透明度研究[D]. 山东大学

[20] 黄伯荣,廖序东. 现代汉语（增订第五版）[M].高等教育出版社.2011

chinaXiv:202011.00125v1